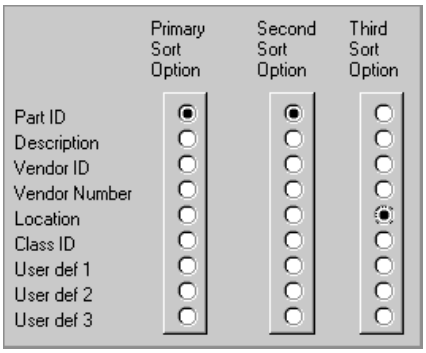
IV - Avaliação

IV.4 – Análise de Dados da Avaliação

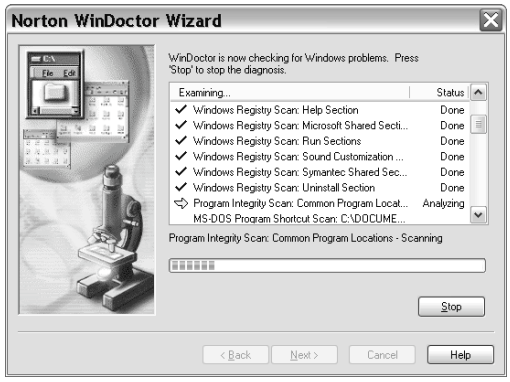
Interactive System Design, Cap. 10,

William Newman

Melhor e Pior?



Melhor e Pior?



Resumo Aula Anterior

- Avaliação com utilizadores
  - Local (Laboratório, Ambiente de trabalho)
  - Utilizadores
  - Métodos de avaliação
    - Observação
    - Interrogação e Entrevista
    - Monitorização fisiológica
  - Vídeos de exemplos de avaliação

Sumário

- Testes com utilizadores
- Fases do Teste
- Análise dos dados por métodos estatísticos
  - Teste t
  - Intervalos de Confiança
  - Chi Quadrado

Antes dos Testes (Planear)

- O plano de testes deve ser definido antes de começar os testes:
  - Objectivo do teste: O que se pretende atingir?
  - Onde e quando serão os testes?
  - Quanto durará cada sessão de testes?
  - Qual o equipamento necessário?
  - Qual o software que é necessário para o teste?

Antes dos Testes

- Qual deve ser o estado do sistema no início do teste?
- Qual deve ser a carga e tempo de resposta do sistema?
- Quem irá conduzir os testes?
- Quem serão os utilizadores que farão os testes e como os arranjaremos?
- Quantos utilizadores serão necessários?
- Que tarefas serão pedidas aos utilizadores?

Antes dos Testes

- Que critério será usado para determinar o fim da execução de uma tarefa correctamente?
- Que ajudas (manuais, ajuda online, etc.) estarão disponíveis ao utilizador durante os testes?
- Em que medida se pode ajudar os utilizadores durante os testes?
- Que dados irão ser recolhidos e como serão analisados?
- Qual será o critério que dirá que a interface é um sucesso?

Testes Piloto

- Não fazer testes sem antes experimentar os procedimentos com 1 ou 2 sujeitos piloto
  - Qualquer pessoa, não precisa de pertencer aos potenciais utilizadores
  - Instruções para os testes são compreensíveis?
  - Questionários?
  - O tempo planeado para cada tarefa é realista?
  - Refinar o procedimento experimental e clarificar aquilo que se vai medir (ex. O que é um erro)

Ex. de Protocolo Experimental

TITLE: Protocol for user evaluation of the Sketch-Based Retrieval prototype at CENTIMFE
AUTHOR: Manuel João da Fonseca, Alfredo Ferreira Junior
DATE: 3 June 2004
PAGES: 10
VERSION: 0.2
KEYWORDS:
CIRCULATION: Consortium

Introduction

This document presents the protocol specified for user evaluation tests to perform at CENTIMFE, to evaluate our Sketch-Based Retrieval prototype.

Our experimental tests will involve eight users from CENTIMFE and Hamibal Abrantes. We divide the experiment in three parts, as described in this protocol. First, we will give a brief description of the experiment, then users will perform a set of tasks and finally they will answer a questionnaire.

With this experimental evaluation we want to get some feedback from users about our second Sketch-Based Retrieval prototype, about returned results and about the overall retrieving process. We will videotape all experiments, measuring times and taking notes of users' comments or suggestions.

Measures

During the experiment we plan to collect the following information:

- Videotape the experiment
- Number of Undos
- Number of Deletes
- Number of Selects & Queries
- Number of News (new query) per query
- Number of queries per search until the final results
- Time spent before finding the desired result
- Number of strokes sketched per query

Fases do Teste

1. Preparação
2. Introdução
3. Realização do Teste
4. Debriefing

Preparação

- Durante a preparação da experiência, o coordenador deve assegurar que:
  - A sala de testes está pronta
  - O sistema está no estado planeado
  - Todo o material para testes está disponível (instruções, questionários)
  - Preparar tudo antes da chegada do utilizador
    - Desligar screensavers e outros programas que podem interromper (Msn, Skype, email, etc.)

Introdução

- O coordenador dos testes
  - Dá as boas vindas ao utilizador
  - Explica brevemente o propósito dos testes
  - Explica o procedimento dos testes
  - Se tiver algum formulário de consentimento (para filmar, fotografar, etc.) deve ser apresentado na introdução.
    - Este deve ser curto e usar linguagem simples e acessível ao utilizador

Introdução – Elementos a Referir

- O objectivo é avaliar o sistema e não o utilizador
- Utilizador pode falar livremente sem medo de ferir susceptibilidades
- Os resultados do teste serão usados para melhorar a interface
- Explicar se vão gravar audio e vídeo
- Referir que o utilizador pode fazer as perguntas que quiser, mas não durante o teste.
- Instruções específicas para a experiência a realizar (ex. Pensar em voz alta, ou realizar a tarefa tão rápido quanto possível sem ajuda)

Durante a Realização do Teste

O coordenador

- Não deve
  - Interagir com o utilizador
  - Fazer comentários
  - Ajudar o utilizador
- Deve
  - Tomar nota das medidas de desempenho
  - Tomar nota dos comentários e observações do utilizador

Medidas de Desempenho

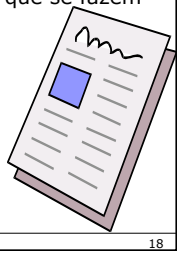
- Medidas de Usabilidade típicas
  - Tempo para completar uma tarefa
  - Nº de tarefas concluídas num determinado intervalo de tempo
  - Nº de erros cometidos
  - Tempo gasto a recuperar de erros
  - Relação entre interacções com sucesso e erros
  - Nº comandos ou funções usadas pelo utilizador
  - Nº comandos ou funções que nunca foram usadas
  - Nº de funções do sistema que o utilizador consegue recordar no debriefing
  - Frequência de utilização de manuais ou ajudas
  - Etc.

Debriefing

- Pedir ao utilizador para preencher um questionário de satisfação
  - Antes de qualquer discussão sobre o sistema
- Pedir comentários sobre o sistema
  - Recebem-se comentários muito diferentes
  - Mas, por vezes servem para um novo desenho
- Coordenador do teste deve (depois do utilizador sair)
  - Garantir que toda a informação recolhida está identificada com o utilizador
  - Escrever um pequeno relatório sobre a experiência (enquanto tudo está fresco)

Relatório Avaliação

- Relatório contendo:
  - Objectivos
  - Descrição do sistema a testar
  - Breve descrição do ambiente em que se fazem as tarefas
  - Características dos participantes
  - Metodologia
  - Tarefas
  - Testes e medidas
  - Análise dos dados medidos



Estatística

Interactive System Design, Cap. 10,

William Newman

Grandezas Estatísticas

$$\bar{X} = \frac{\sum X_i}{N}$$

$$SQ = \sum (X_i - \bar{X})^2$$

$$df = N - 1$$

$$s^2 = \frac{SQ}{N-1}$$

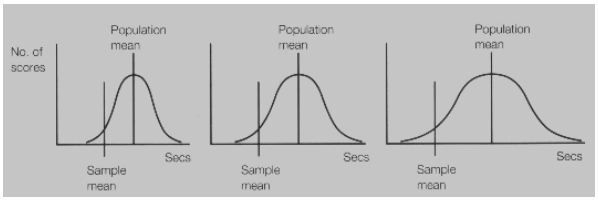
$$s = \sqrt{s^2}$$

- Média
- Soma dos quadrados das diferenças
- Graus de liberdade
- Variância
- Desvio padrão

Exemplo de Teste

- Métrica de desempenho: Execução <=30min
- Teste com 6 utilizadores
  - Teste dá: 20, 15, 40, 90, 10, 5
  - Média = 30
  - Desvio padrão = 32
  - Parece ok!?
  - Errado, nada se pode afirmar
- Factores que contribuem para a incerteza?
  - Pequeno nº de utilizadores no teste (n=6)
  - Resultados muito variáveis (desvio padrão =32)
    - Desvio padrão = dispersão do valor médio (-2;62)

Significado Variância/Desvio Padrão



Testes Estatísticos

- Experimentação Controlada
  - Responder a:
    - Solução A melhor que a B?
      - Tendo em conta a sua usabilidade.
      - Exemplos:
        - » Alteração no tipo de menus (PullDown vs Pie)
        - » Caligráfico vs Menus
    - Solução cumpre os objectivos?
      - Os nossos objectivos de usabilidade são atingidos

Testes Estatísticos

- Procedimento
  - Escolha da população "significativa"
    - Público alvo
  - Formulação da hipótese nula (H_0)
  - Realização dos testes
    - Recolher os dados
  - Conclusão
    - Analisar os resultados
    - Aplicar tratamento estatístico

Hipótese Nula e Grau de Confiança

- Hipótese H_0 – Hipótese Nula
  - Diz o contrário do que pretendemos
    - Ex. A mudança de menus não afecta o desempenho
- Hipótese H_1 - Hipótese experimental
  - Diz o que queremos verificar
    - Ex. Os novos menus melhoram o desempenho
- Objectivo: rejeitar H_0 e aceitar H_1
  - Demonstramos que H_0 é falsa para um determinado α (valor típico 0,05)

Comparar duas Alternativas

- Experiência ***entre grupos***
  - Dois grupos de teste
  - Cada grupo usa *apenas* um dos sistemas
- Experiência **Intragrupos**
  - Um grupo de utilizadores
    - Cada pessoa usa *ambos* os sistemas
    - Não podem usar as mesmas tarefas ou pela mesma ordem (aprendizagem)
  - Melhor para técnicas de interacção básicas
- Entre grupos requer mais participantes
- Ver se as diferenças são estatisticamente significativas
  - Assume distribuição normal & mesmo desvio padrão

Comparar 2 amostras – Teste de t

Objectivo : qual das duas é melhor

- Variância combinada
- Desvio padrão da diferença
- Valor de t

$$s^2 = \frac{(SQ_1 + SQ_2)}{N_1 + N_2 - 2}$$

$$s_{ed} = \sqrt{s^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

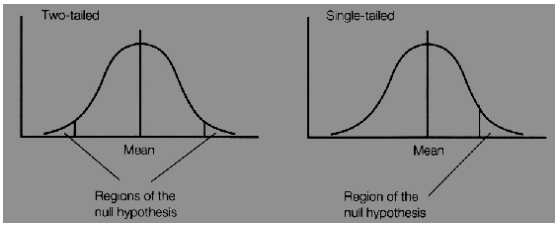
$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{ed}}$$

- Se t > tH_0 (da tabela)
  - então H_0 é falsa (para α)

Tabela de t

Degrees of freedom	$\alpha = 0.10$ $\alpha = 0.05$	$\alpha = 0.05$ $\alpha = 0.025$	$\alpha = 0.02$ $\alpha = 0.01$	$\alpha = 0.01$ $\alpha = 0.005$	(two-tailed) (single-tailed)
1	6.314	12.706	31.821	63.656	
2	2.920	4.303	6.965	9.925	
3	2.353	3.182	4.541	5.841	
4	2.132	2.776	3.747	4.604	
5	2.015	2.571	3.365	4.032	
6	1.943	2.447	3.143	3.707	
7	1.895	2.365	2.998	3.499	
8	1.860	2.306	2.896	3.355	
9	1.833	2.262	2.821	3.250	
10	1.812	2.228	2.764	3.169	
11	1.796	2.201	2.718	3.106	
12	1.782	2.179	2.681	3.055	

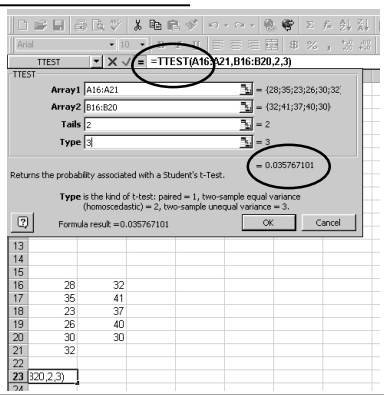
Bicaudal e Unicaudal



Exemplo: Teste de t - Bilheteira

- Objectivo: Comparar Bilheteira com Máquina
- Hipótese nula:
  - a forma de aquisição do bilhete não tem influência no tempo da tarefa
- Medidas
  - bilheteira: 28, 25, 23, 26, 30, 32 segundos
  - máquina: 32, 41, 37, 40, 30 segundos
- Médias
  - bilheteira: 29 s máquina: 36 s

Teste de t - Bilheteira



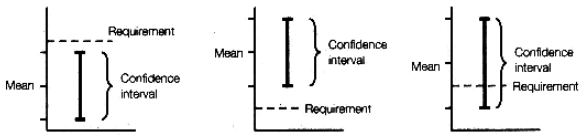
Teste de t - Bilheteira

- Constata-se que:
  - as duas amostras têm uma probabilidade de (apenas) 3,6% de serem a mesma amostra
  - Rejeita-se H_0 , pois 0.036 < 0.05 (significância p)
- Conclusão
  - A compra de bilhetes em máquina é 24% (36/29) mais lenta com uma probabilidade de 96,4%

Intervalo de Confiança

- Testar uma amostra contra um valor limite
- Intervalo de confiança
  - 2 extremos entre os quais a média da população está compreendida com uma dada probabilidade
- Exemplo
  - uma operação não deve demorar mais do que 25s -> intervalo totalmente abaixo de 25 s

Intervalo de Confiança



Evitar este

Intervalo de Confiança

- Calcular variância (S^2)
- Desvio padrão da média $s_{em} = \sqrt{S^2 / N}$
- Determinar t unicaudal para a probabilidade pretendida e grau de liberdade da amostra
- O intervalo estará compreendido entre

$$X_{\min} = \overline{X} - \left(t_{p,gl} \times s_{em}\right)$$

$$X_{\max} = \overline{X} + \left(t_{p,gl} \times s_{em}\right)$$

Intervalo de Confiança (Ex.)

•Métrica: Nº de Erros
•Objectivo: nº erros <= 15
•Amostra: 13, 6, 8, 11
•Média: 9,5
•Variância: 9,67

- Desvio padrão da média: $s_{em} = \sqrt{9,67/4} = 1,55$
- H_0 – Nº de erros superior a 15
- Para p=0,05 t=3,182 (da tabela ou usando TINV ())
- Intervalo
  - x_{min} = 9,5 – 3,182 x 1,55 = 4,54
  - x_{max} = 9,5 + 3,182 x 1,55 = 14,43
- Intervalo abaixo de 15
  - Rejeitar H_0 => Nº erros inferior a 15 c/ 95% de certeza

Teste do Chi Quadrado

- Dados correspondentes a uma ou mais categorias
  - Ex: determinar preferência entre várias opções de escolha
- Procedimento:
  - cálculo da diferença entre as frequências observadas e as esperadas

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Teste do Chi Quadrado (Exemplo)

- Qual a opção preferida de entre as 3?
- H_0 – Preferência igual pelas 3
- 30 utilizadores

Opção	f esperada	f observada	Diferença	Quad. Da Diferença	f/f esperada
A	10	7	-3	9	0,9
B	10	13	3	9	1,3
C	10	10	0	0	1

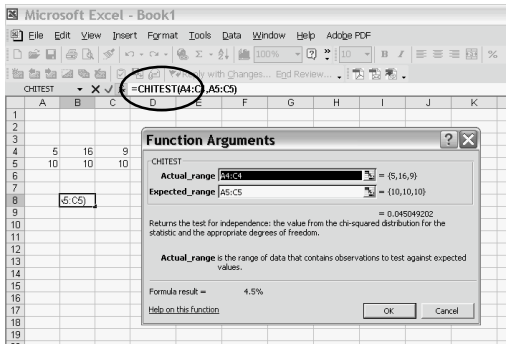
$\chi^2 = 6.2$

- Graus de liberdade: N=3-1=2
- Da tabela obtemos 5,99 para p=0,05
- Rejeita-se a hipótese nula (5.99 < 6.2)

Tabela do chi Quadrado

Degrees of freedom	$\alpha = 0.05$	$\alpha = 0.01$
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21

Chi Quadrado no Excel



- 0.045 < 0.05 -> Rejeitar H_0

Escolha de métodos ou algoritmos?

- Todas as funções estatísticas estão disponíveis em bibliotecas:
  - Matlab, SPSS, ou mesmo Excel
- Assim o que é preciso é **saber escolher o método** mais apropriado

Resumo

- Os testes mais conclusivos devem ser realizados com utilizadores reais
- Os testes devem ser planeados e aprovados previamente
- Devem ser recolhidos dados qualitativos e quantitativos
- Dados numéricos só são conclusivos se validados por testes estatísticos
- Escolhido o método estatístico adequado
  - Usar as ferramentas existentes

Próxima Aula

- Documentação e Ajudas
- Importância, características e guia de estilos dos manuais
- Manuais convencionais
- Princípios e guias para manuais
- Ajudas Interactivas

- Ler: HCI, Cap. 11, Alan Dix